

Evaluation of Generalized Degrees of Freedom for Sparse Estimation by Replica Method

A Sakata

The Institute of Statistical Mathematics, Midori-cho, Tachikawa 190-8562, Japan
The Graduate University for Advanced Science (SOKENDAI), Hayama-cho,
Kanagawa 240-0193, Japan

E-mail: ayaka@ism.ac.jp

Abstract. We develop a method to evaluate the generalized degrees of freedom (GDF) for linear regression with sparse regularization. The GDF is a key factor in model selection, and thus its evaluation is useful in many modelling applications. An analytical expression for the GDF is derived using the replica method in the large-system-size limit with random Gaussian predictors. The resulting formula has a universal form that is independent of the type of regularization, providing us with a simple interpretation. Within the framework of replica symmetric (RS) analysis, GDF has a physical meaning as the effective fraction of non-zero components. The validity of our method in the RS phase is supported by the consistency of our results with previous mathematical results. The analytical results in the RS phase are calculated numerically using the belief propagation algorithm.

Keywords: Cavity and replica method, Statistical inference, Learning theory

Submitted to: *J. Stat. Mech.*

1. Introduction

Statistical modelling plays a key role in extracting the structures of a system that may be hidden behind observed data and using them for prediction or control. A statistical model approximates the true generative process of the data, which is generally expressed by a probability distribution. Although it is necessary to adopt an appropriate statistical model, this will depend on the purpose of the modelling, and the definition of appropriateness is not unique. Akaike proposed an information criterion for model selection, where the appropriate model is defined using Kullback–Leibler divergence [1]. This criterion validates the relative effectiveness of the model under consideration, and mathematically expresses the contribution of the model to the prediction performance.

Since the systemization of the least absolute shrinkage and selection operator (LASSO) [2], which simultaneously achieves variable selection and estimation, sparse estimation has been attracting considerable attention in fields such as signal processing

[3, 4] and machine learning [5, 6]. In general, sparse estimation is formulated as the problem of minimizing the estimating function penalized by sparse regularization. The estimated variables have zero components, a property known as sparsity. To find the sparse representation of the system from among various candidates, a seemingly hidden rule that controls the system is sought. Similar to LASSO, ℓ_1 regularization is widely used because of its convexity, which yields mathematical and algorithmic tractability [3]. In addition, non-convex regularization, such as using the ℓ_p ($p < 1$)-norm [7, 8], has been studied to obtain a sparser representation than that given by ℓ_1 -norm regularization [9]. Furthermore, the smoothly clipped absolute deviation (SCAD) and adaptive LASSO penalty have been investigated [10, 11, 12] to acquire the oracle property, which the LASSO estimator does not possess.

The emergence of the estimation paradigm associated with sparsity requires the development of appropriate model selection criteria. In sparse estimation, the determination of the regularization parameter can be regarded as the selection of a model from a family of models that have different sparsities controlled by the regularization parameter. In addition to the cross-validation (CV) method [13], which is a simple numerical approach for sparse estimation [14, 15], analytical model selection methods with lower computational costs have been developed. One such method involves estimating the *generalized degrees of freedom* (GDF) [16]. The GDF is a key quantity for Mallows' C_p , a model selection criterion based on the prediction error [17]. In particular, the derivation of GDF has been studied in linear regression with a known variance [18]. The analytical form of the GDF for LASSO [19] and elastic net regularization [20] are well known, but general expressions for other regularizations have not yet been derived.

In this paper, we propose an analytical method based on statistical physics for the derivation of GDF in sparse estimation. Certain aspects of statistical physics developed for random systems have already been applied to sparse estimation problems [21, 22, 23, 24]. The analysis of typical properties provides physical interpretations of the problems based on phase transition pictures, and this contributes to the development of algorithms [25, 26, 27, 28]. The statistical physical method can be applied to the estimation of GDF for sparse regularization. We show that GDF is expressed as the effective fraction of non-zero components for any sparse regularization. This expression is a mathematical realization of the meaning of GDF in terms of “model complexity” [19, 29].

The remainder of this paper is organized as follows. Section 2 summarizes the model selection criterion discussed in this paper and highlights some previous related studies on sparse estimation. Section 3 explains our problem setting for the estimation of GDF. Sections 4 and 5 describe our analytical method based on the replica method for sparse estimation. Section 6 represents the behaviour of GDF for ℓ_1 , elastic net, ℓ_0 , and SCAD regularization. Section 7 proposes the numerical calculation of GDF using the belief propagation algorithm, and discusses the generality of the results. In Section 8, the approximation performance of our method for the calculation of GDF is examined in the case of ℓ_0 regularization. Finally, Section 9 concludes the paper.

2. Overview of model selection

In this section, we explain the criteria for model selection discussed in this paper. In addition, we summarize previous studies and identify our contributions. We focus on the parametric model, where the true generative model of z , denoted by $q(z)$, is approximated by $p(z|\boldsymbol{\theta})$ with a parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^N$, where $\boldsymbol{\Theta}$ is a parameter space. The parameter is estimated under the given model to effectively describe the true distribution using training data $\boldsymbol{w} = \{w_\mu\}$ ($\mu = 1, \dots, M, w_\mu \sim q(w_\mu)$). Let us prepare a set of candidate models $\mathcal{M} = \{p_1(z|\hat{\boldsymbol{\theta}}_1), \dots, p_m(z|\hat{\boldsymbol{\theta}}_m)\}$ for the approximation of the true distribution, where $\hat{\boldsymbol{\theta}}_k$ is the estimated parameter under the k -th model. Model selection is then the problem of adopting a model based on a certain criterion.

2.1. Information criterion

The information criterion evaluates the quality of the statistical model based on Kullback–Leibler (KL) divergence. KL divergence describes the closeness between the true distribution $q(z)$ and the assumed distribution $p(z|\hat{\boldsymbol{\theta}}_{\text{ML}}(\boldsymbol{w}))$ as

$$\text{KL}(q : p) = E_{z \sim q(z)} [\log q(z)] - E_{z \sim q(z)} [\log p(z|\hat{\boldsymbol{\theta}}_{\text{ML}}(\boldsymbol{w}))], \quad (1)$$

where $\hat{\boldsymbol{\theta}}_{\text{ML}}(\boldsymbol{w})$ is the maximum likelihood estimator from the training sample \boldsymbol{w} . The dependency on the model appears only in the second term of (1), called the predicting log-likelihood, i.e. $l(\boldsymbol{w}) \equiv E_{z \sim q(z)} [\log p(z|\hat{\boldsymbol{\theta}}_{\text{ML}}(\boldsymbol{w}))]$. Therefore, the maximization of the predicting log-likelihood is the basis for the information criterion. Unfortunately, it is generally impossible to evaluate the predicting log-likelihood, because we cannot determine the true distribution. We define the estimator of the predicting log-likelihood using the empirical distribution

$$\hat{l}(\boldsymbol{w}) = \frac{1}{M} \sum_{\mu=1}^M \log p(w_\mu|\hat{\boldsymbol{\theta}}_{\text{ML}}(\boldsymbol{w})), \quad (2)$$

which corresponds to the maximum log-likelihood. The expected value of the difference between the predicting log-likelihood and the maximum log-likelihood, termed the bias, is given by

$$b = E_{\boldsymbol{w} \sim q(\boldsymbol{w})} \left[\hat{l}(\boldsymbol{w}) - E_{z \sim q(z)} [\log p(z|\hat{\boldsymbol{\theta}}_{\text{ML}}(\boldsymbol{w}))] \right], \quad (3)$$

where $q(\boldsymbol{w}) = \prod_{\mu} q(w_\mu)$. The information criterion is defined as an unbiased estimator of the negative predicting log-likelihood:

$$\text{IC}(\boldsymbol{w}) = -2\hat{l}(\boldsymbol{w}) + 2\hat{b}(\boldsymbol{w}), \quad (4)$$

where $\hat{b}(\boldsymbol{w})$ is an unbiased estimator of the bias, and the coefficient 2 is a conventional value. The optimal model is defined as that which minimizes $\text{IC}(\boldsymbol{w})$ among the models in \mathcal{M} . Intuitively, the first and second terms represent the training error and the complexity of the model, respectively. As the complexity of the model increases, the model can express various distributions. However, overfitting is likely to occur, which hampers the prediction of unknown data. The information criterion selects the model

that achieves the best trade-off between the training error and the level of model complexity.

The values of b can be calculated asymptotically. In particular, when the statistical model contains the true model, namely a parameter $\boldsymbol{\theta}^*$ exists such that $q(z) = p(z|\boldsymbol{\theta}^*)$, the information criterion is known as Akaike's information criterion (AIC), where the bias term b is reduced to the dimension of the parameter $\boldsymbol{\theta}$ [1].

The criterion explained thus far is for models constructed by maximum likelihood estimation. To determine the parameter with other learning strategies, we focus on maximum likelihood estimation under regularization, where the GDF facilitates the extension of the information criterion [29]. A general expression of GDF is naturally derived from another model selection criterion, namely, Mallows' C_p [17].

2.2. Mallows' C_p and generalized degrees of freedom

The prediction of unknown data is another criterion for the evaluation of a model. We define the squared prediction error per component as

$$\text{err}_{\text{pre}}(\mathbf{w}) = \frac{1}{M} E_{\mathbf{z}} [\|\mathbf{z} - \hat{\mathbf{w}}(\mathbf{w})\|_2^2], \quad (5)$$

where $\hat{\mathbf{w}}$ is the estimate of \mathbf{w} and $\mathbf{z} \in \mathbb{R}^M$ is independent of $\mathbf{w} \in \mathbb{R}^M$, but each component of \mathbf{z} is generated according to the same distribution as \mathbf{w} . When the training sample is generated as $\mathbf{w} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_M)$, where \mathbf{I}_M is the M -dimensional identity matrix, Mallows' C_p , calculated as

$$c_p(\mathbf{w}) = \text{err}_{\text{train}}(\mathbf{w}) + 2\sigma^2 \hat{\text{df}}(\mathbf{w}), \quad (6)$$

is an unbiased estimator of the prediction error. Here,

$$\text{err}_{\text{train}}(\mathbf{w}) = \frac{1}{M} \|\mathbf{w} - \hat{\mathbf{w}}(\mathbf{w})\|_2^2 \quad (7)$$

is the training error and $\hat{\text{df}}(\mathbf{w})$ is an unbiased estimator of GDF defined by

$$\text{df} = \frac{\text{cov}(\mathbf{w}, \hat{\mathbf{w}}(\mathbf{w}))}{M\sigma^2}, \quad (8)$$

which quantifies the complexity of the model [19, 29], where $\text{cov}(\mathbf{w}, \hat{\mathbf{w}}(\mathbf{w})) = E_{\mathbf{w}}[(\mathbf{w} - E_{\mathbf{w}}[\mathbf{w}])(\hat{\mathbf{w}}(\mathbf{w}) - E_{\mathbf{w}}[\hat{\mathbf{w}}(\mathbf{w})])]$. In the framework of C_p , the optimal model is defined as that which minimizes $c_p(\mathbf{w})$ among the models in \mathcal{M} .

Another expression of GDF is given by [16]

$$\text{df} = \frac{1}{M} \sum_{\mu} E_{\mathbf{w}} \left[\frac{\partial \hat{w}_{\mu}(\mathbf{w})}{\partial w_{\mu}} \right], \quad (9)$$

which corresponds to the expectation of Stein's unbiased risk estimate (SURE) for the prediction error [30]. GDF was originally introduced as an extension of the degrees of freedom in the linear estimation rule for a general modelling procedure in the form (9) [16].

When the assumed model obeys a Gaussian distribution $p(\mathbf{w}|\boldsymbol{\theta}) \propto \exp(-\frac{1}{2\sigma^2}||\mathbf{w} - \boldsymbol{\mu}(\boldsymbol{\theta})||_2^2)$ with a known variance, and taking $\boldsymbol{\mu}(\hat{\boldsymbol{\theta}}_{\text{ML}}(\mathbf{w})) = \hat{\mathbf{w}}(\mathbf{w})$, AIC (normalized by the number of training samples) is given by [19]

$$\text{AIC}(\mathbf{w}) = \frac{\text{err}_{\text{train}}(\mathbf{w})}{\sigma^2} + 2\hat{\text{df}}(\mathbf{w}). \quad (10)$$

Equations (6) and (10) indicate that model selection based on AIC and that based on C_p give the same result; they are proportional to each other $c_p(\mathbf{w}) = \sigma^2 \text{AIC}(\mathbf{w})$.

2.3. Model selection for sparse regularization and our contributions

The regression problems with sparse regularization is formulad as

$$\min_{\mathbf{x}} \{e(\mathbf{x}; \mathbf{w}, \mathbf{A}) + r(\mathbf{x}; \eta)\}, \quad (11)$$

where $e(\mathbf{x}; \mathbf{w}, \mathbf{A})$ measures the difference between training data \mathbf{w} and its fit using regression coefficients \mathbf{x} under the predictor matrix \mathbf{A} , and $r(\mathbf{x}; \eta)$ is the regularization term with the regularization parameter η that enhances zero components in \mathbf{x} . The regularization parameter determines the number of predictors used in the expression of the data distribution, and the model distribution under the determined number of predictors can be regarded as a model: $\mathcal{M} = \{p_\eta(z|\hat{\boldsymbol{\theta}}_\eta)|\eta \in \mathbf{H}\}$, where \mathbf{H} is the support of the regularization parameter. Therefore, tuning the regularization parameter η corresponds to model selection. However, in general, the derivation of AIC based on the asymptotic expansion is not straightforwardly applicable to sparse regularization. In such cases, C_p is useful for deriving the model selection criterion when the squared error is considered. In LASSO, it is mathematically proven that, when the number of training samples is greater than the number of predictors, the ratio of the number of non-zero regression coefficients to the number of training samples is an unbiased estimator of the degrees of freedom in a finite sample [19]. However, the derivation of GDF is analytically difficult for general sparse regularizations. To overcome this difficulty, GDF computation techniques have been developed using the parametric bootstrap method [18] and SURE [30, 31].

In the present paper, we propose an estimation technique for GDF using the replica method under a replica symmetric (RS) assumption for linear regression with Gaussian i.i.d. predictors. The replica symmetric analysis for the estimation problems under sparse regularization are shown in [21, 22, 23, 24]. In these papers, the replica method is employed to study phase transition or the property of estimators. We extend this analytical method for the calculation of GDF that is not taken into account in the current formalism of the replica analysis. The technique we propose is applicable to general sparse regularization. Using our method, the correspondence between GDF and the effective fraction of non-zero components in the large-system-size limit is shown to be independent of the form of regularization. Our approach differs from previous methods in which GDF has been derived for specific types of regularization. We apply our method to ℓ_1 , elastic net, ℓ_0 , and SCAD regularization to obtain the GDF. The results shown here

for ℓ_1 and elastic net regularization are weaker than those in previous studies, where the unbiased estimator of GDF, $\hat{\text{df}}$, is derived for one instance of the predictor. However, our method is consistent with previous results, which supports the validity of our approach. Furthermore, our method can be applied to non-convex sparse regularizations such as ℓ_0 and SCAD, and extends the discussion of GDF to general sparse regularization. For the ℓ_0 case, the solution under the RS assumption is always unstable against perturbations that break the replica symmetry, but we show that GDF under the RS assumption approximates the true value of GDF. In the case of SCAD regularization, our method can identify the most appropriate model based on the prediction error within the range of the RS assumption when the mean of the data is sufficiently small. The generality of the result in terms of the correspondence between GDF and the effective fraction of non-zero components is discussed using a belief propagation algorithm for other predictor matrices.

3. Problem setting and formulation

We apply a linear regression model with sparse regularization $r(\mathbf{x}; \eta) = \sum_i r(x_i; \eta)$, where η is a regularization parameter, to a set of training data $\mathbf{y} \in \mathbb{R}^M$:

$$\min_{\mathbf{x}} \left\{ \frac{1}{2} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + r(\mathbf{x}; \eta) \right\}, \quad (12)$$

where the column vectors of $\mathbf{A} = \{\mathbf{A}_1, \dots, \mathbf{A}_N\} \in \mathbb{R}^{M \times N}$ and components of $\mathbf{x} \in \mathbb{R}^N$ correspond to predictors and regression coefficients, respectively. Here, the coefficient of the squared error, $1/2$, is introduced for mathematical convenience. The variable \mathbf{x} to be estimated here corresponds to the parameter $\boldsymbol{\theta}$ in the previous section, and the number of non-zero components in \mathbf{x} corresponds to the number of parameters used in the model. We introduce the posterior distribution of \mathbf{x} :

$$P_\beta(\mathbf{x}|\mathbf{y}, \mathbf{A}) = \exp \left\{ -\frac{\beta}{2} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 - \beta r(\mathbf{x}; \eta) - \ln Z_\beta(\mathbf{y}, \mathbf{A}) \right\}, \quad (13)$$

where $Z_\beta(\mathbf{y}, \mathbf{A})$ is the normalization constant. The distribution as $\beta \rightarrow \infty$ is the uniform distribution over the minimizers of (12). Estimate of the solution of (12) under a fixed set of $\{\mathbf{y}, \mathbf{A}\}$, denoted by $\hat{\mathbf{x}}(\mathbf{y}, \mathbf{A})$, is given by

$$\hat{\mathbf{x}}(\mathbf{y}, \mathbf{A}) = \lim_{\beta \rightarrow \infty} \langle \mathbf{x} \rangle_\beta, \quad (14)$$

where $\langle \cdot \rangle_\beta$ denotes the expectation according to (13) at β . Using this estimate $\hat{\mathbf{x}}(\mathbf{y}, \mathbf{A})$ of \mathbf{x} , the training sample \mathbf{y} is estimated as

$$\hat{\mathbf{y}}(\mathbf{y}, \mathbf{A}) = \mathbf{A}\hat{\mathbf{x}}(\mathbf{y}, \mathbf{A}). \quad (15)$$

To understand the typical performance of (12), we calculate the expectation of the training error with respect to \mathbf{y} and \mathbf{A} ,

$$\overline{\text{err}}_{\text{train}} = E_{\mathbf{y}, \mathbf{A}}[\text{err}_{\text{train}}(\mathbf{y}, \mathbf{A})]. \quad (16)$$

At a sufficiently large system size $N \rightarrow \infty$, we set the scaling relationship as $\alpha = M/N \sim O(1)$ and $\hat{\rho} = K/N \sim O(1)$, where K is the number of non-zero components of $\hat{\mathbf{x}}$. The training error relates to the free energy density f as [34]

$$f \equiv - \lim_{\beta \rightarrow \infty} \frac{1}{N} \frac{\partial}{\partial \beta} \phi_\beta = \frac{\alpha \overline{\text{err}}_{\text{train}}}{2} + \bar{r}, \quad (17)$$

where $\phi_\beta \equiv E_{\mathbf{y}, \mathbf{A}}[\ln Z_\beta(\mathbf{y}, \mathbf{A})]$ and

$$\bar{r} = \frac{1}{N} E_{\mathbf{y}, \mathbf{A}}[r(\hat{\mathbf{x}}(\mathbf{y}, \mathbf{A}); \eta)]. \quad (18)$$

The expectation of the regularization term \bar{r} is derived separately from f , as shown in the following section. Hence, the training error is derived as

$$\overline{\text{err}}_{\text{train}} = \frac{2(f - \bar{r})}{\alpha}. \quad (19)$$

For the calculation of GDF, we introduce external fields κ and ν , and define the extended posterior distribution as

$$P_{\beta, \kappa, \nu}(\mathbf{x} | \mathbf{y}, \mathbf{A}) = \exp \left\{ -\frac{\beta}{2} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 - \beta r(\mathbf{x}; \eta) - \beta \sum_{\mu i} (\kappa y_\mu + \nu) A_{\mu i} x_i - \ln Z_{\beta, \kappa, \nu}(\mathbf{y}, \mathbf{A}) \right\}, \quad (20)$$

where $Z_{\beta, \kappa, \nu}(\mathbf{y}, \mathbf{A})$ is the normalization constant. We define the extended free energy density as

$$f_{\kappa, \nu} = - \lim_{\beta \rightarrow \infty} \frac{1}{N} \frac{\partial}{\partial \beta} \phi_{\beta, \kappa, \nu}, \quad (21)$$

where $\phi_{\beta, \kappa, \nu} = E_{\mathbf{y}, \mathbf{A}}[\ln Z_{\beta, \kappa, \nu}(\mathbf{A}, \mathbf{y})]$ and $f = f_{\kappa=0, \nu=0}$. We derive the following quantities from the extended free energy density:

$$\hat{\gamma} \equiv \frac{1}{M} \sum_{\mu} E_{\mathbf{y}, \mathbf{A}}[y_{\mu} \sum_i A_{\mu i} \hat{x}_i(\mathbf{y}, \mathbf{A})] = \frac{1}{\alpha} \frac{\partial}{\partial \kappa} f_{\kappa, \nu} \Big|_{\kappa, \nu=0} \quad (22)$$

$$\hat{m}_y \equiv \frac{1}{M} \sum_{\mu} E_{\mathbf{y}, \mathbf{A}}[\sum_i A_{\mu i} \hat{x}_i(\mathbf{y}, \mathbf{A})] = \frac{1}{\alpha} \frac{\partial}{\partial \nu} f_{\kappa, \nu} \Big|_{\kappa, \nu=0}. \quad (23)$$

Using these, the GDF for a Gaussian training sample is derived as

$$\text{df} = \frac{\hat{\gamma} - \hat{m}_y m_y}{\sigma_y^2}, \quad (24)$$

where $m_y \in \mathbb{R}$ and $\sigma_y^2 \in \mathbb{R}$ are the mean and variance of the training sample, respectively. Further, C_p given by (6) is the unbiased estimator of the prediction error. Hence, the expectation of the prediction error with respect to \mathbf{y} and \mathbf{A} ,

$$\overline{\text{err}}_{\text{pre}} = \frac{1}{M} E_{\mathbf{y}, \mathbf{A}}[E_{\mathbf{z}}[\|\mathbf{z} - \hat{\mathbf{y}}(\mathbf{y}, \mathbf{A})\|_2^2]], \quad (25)$$

is given by

$$\overline{\text{err}}_{\text{pre}} = \overline{\text{err}}_{\text{train}} + 2\sigma_y^2 \text{df}. \quad (26)$$

4. Analysis

For the derivation of GDF, we resort to the replica method [32, 33]. The RS calculations for ℓ_0 and ℓ_1 minimization are shown in the typical performance analysis of compressed sensing [21] and dictionary learning [24]. We summarize the analytical method and explain how it can be extended to the evaluation of GDF. Hereafter, we consider Gaussian i.i.d. predictors $A_{\mu i} \sim \mathcal{N}(0, M^{-1}) \forall (\mu, i)$.

4.1. Replica method and replica symmetry

We calculate the generating function ϕ_β using the following identity:

$$E_{\mathbf{y}, \mathbf{A}}[\ln Z_\beta(\mathbf{y}, \mathbf{A})] = \lim_{n \rightarrow 0} \frac{E_{\mathbf{y}, \mathbf{A}}[Z_\beta^n(\mathbf{y}, \mathbf{A})] - 1}{n}. \quad (27)$$

Assuming that n is a positive integer, we can express the expectation of $Z_\beta^n(\mathbf{y}, \mathbf{A})$ by introducing n replicated systems:

$$\begin{aligned} E_{\mathbf{y}, \mathbf{A}}[Z_\beta^n(\mathbf{y}, \mathbf{A})] &= \int d\mathbf{A} d\mathbf{y} P_A(\mathbf{A}) P_y(\mathbf{y}) \int d\mathbf{x}^{(1)} \dots d\mathbf{x}^{(n)} \\ &\quad \times \exp \left[\sum_{a=1}^n \left\{ -\frac{\beta}{2} \|\mathbf{y} - \mathbf{A} \mathbf{x}^{(a)}\|_2^2 - \beta r(\mathbf{x}^{(a)}; \eta) \right\} \right], \end{aligned} \quad (28)$$

where $P_A(\mathbf{A}) = \prod_{\mu, i} \sqrt{\frac{M}{2\pi}} \exp(-\frac{M}{2} A_{\mu i}^2)$ and $P_y(\mathbf{y}) = \prod_{\mu} \sqrt{\frac{1}{2\pi\sigma_y^2}} \exp(-\frac{1}{2\sigma_y^2} (y - m_y)^2)$. We characterize the microscopic states of $\{\mathbf{x}^{(a)}\}$ with the macroscopic quantities

$$q^{(ab)} = \frac{1}{M} \sum_i x_i^{(a)} x_i^{(b)}. \quad (29)$$

Introducing the identity for all combinations of a, b ($a \leq b$)

$$1 = \int dq^{(ab)} \delta \left(q^{(ab)} - \frac{1}{M} \sum_i x_i^{(a)} x_i^{(b)} \right), \quad (30)$$

the integration with respect to \mathbf{A} leads to the following expression:

$$\begin{aligned} E_{\mathbf{y}, \mathbf{A}}[Z_\beta^n(\mathbf{y}, \mathbf{A})] &= \int d\mathcal{Q} \mathcal{S}(\mathcal{Q}) \int \{d\mathbf{u}^{(a)}\} P_u(\{\mathbf{u}^{(a)}\} | \mathcal{Q}) \int d\mathbf{y} P_y(\mathbf{y}) \\ &\quad \times \exp \left\{ -\frac{\beta}{2} \sum_a \|\mathbf{y} - \mathbf{u}^{(a)}\|_2^2 \right\}, \end{aligned} \quad (31)$$

where each component of $\mathbf{u}^{(a)}$, denoted by $u_\mu^{(a)}$, is statistically equivalent to $\sum_i A_{\mu i} x_i^{(a)}$, and \mathcal{Q} is a matrix representation of $\{q^{(ab)}\}$. Setting $\tilde{\mathbf{u}}_\mu = \{u_\mu^{(1)}, \dots, u_\mu^{(n)}\}$, its probability distribution is given by [21]

$$P_u(\{\mathbf{u}^a\} | \mathcal{Q}) = \prod_\mu \frac{1}{\sqrt{(2\pi)^n |\mathcal{Q}|}} \exp \left(-\frac{1}{2} \tilde{\mathbf{u}}_\mu^T \mathcal{Q}^{-1} \tilde{\mathbf{u}}_\mu \right), \quad (32)$$

and the function $\mathcal{S}(\mathcal{Q})$ is given by

$$\mathcal{S}(\mathcal{Q}) = \int d\hat{\mathcal{Q}} \{d\mathbf{x}^{(a)}\} \exp \left\{ -M \sum_{a \leq b} q^{(ab)} \hat{q}^{(ab)} + \sum_{a \leq b} \sum_i \hat{q}^{(ab)} x_i^{(a)} x_i^{(b)} \right\}$$

$$- \beta \sum_a r(\mathbf{x}^{(a)}; \eta) \}, \quad (33)$$

where $\hat{q}^{(ab)}$ is the conjugate variable for the integral representation of the delta function in (30), and \hat{Q} is the matrix representation of $\{\hat{q}^{(ab)}\}$.

To obtain an analytic expression with respect to $n \in \mathbb{R}$ and take the limit as $n \rightarrow 0$, we restrict the candidates for the dominant saddle point to those of RS form as

$$(q^{(ab)}, \hat{q}^{(ab)}) = \begin{cases} (Q, -\tilde{Q}/2) & (a = b) \\ (q, \tilde{q}) & (a \neq b). \end{cases} \quad (34)$$

For $\beta \rightarrow \infty$, RS order parameters scale to keep $\beta(Q - q) = \chi$, $\beta^{-1}(\tilde{Q} + \tilde{q}) = \hat{Q}$, and $\beta^{-2}\tilde{q} = \hat{\chi}$ of the order of unity. Under the RS assumption, the free energy density is given by

$$f = \text{extr}_{Q, \chi, \hat{Q}, \hat{\chi}} \left\{ \frac{\alpha(Q + \sigma_y^2 + m_y^2)}{2(1 + \chi)} - \frac{\alpha(Q\hat{Q} - \chi\hat{\chi})}{2} - \frac{1}{2}\pi_r(\hat{Q}, \hat{\chi}) \right\}, \quad (35)$$

where $\text{extr}_{Q, \chi, \hat{Q}, \hat{\chi}}$ denotes extremization with respect to the variables $\{Q, \chi, \hat{Q}, \hat{\chi}\}$. The function π_r , where the subscript r denotes the dependency on the regularization, is given by

$$\pi_r(\hat{Q}, \hat{\chi}) = 2 \int Dz \log g_r(h^{\text{RS}}(z; \hat{\chi}), \hat{Q}) \quad (36)$$

$$g_r(h, \hat{Q}) = \max_x \exp \left(-\frac{\hat{Q}}{2}x^2 + hx - r(x; \eta) \right), \quad (37)$$

where $h^{\text{RS}}(z; \hat{\chi}) = \sqrt{\hat{\chi}}z$ is the random field that effectively represents the randomness of the problem introduced by \mathbf{y} and \mathbf{A} , and $Dz = dz \exp(-z^2/2)/\sqrt{2\pi}$. The solution of x concerned with the effective single-body problem (37), denoted by $x_r^*(z; \hat{Q}, \hat{\chi})$, is statistically equivalent to the solution of the original problem (12). Therefore, the expectation of the regularization term is derived as

$$\bar{r} = \int Dz r(x_r^*(z; \hat{Q}, \hat{\chi}); \eta). \quad (38)$$

The variables $Q, \chi, \hat{Q}, \hat{\chi}$ are determined by saddle point equations to satisfy the extremum conditions of the free energy density:

$$\chi = \frac{1}{\alpha} \frac{\partial \pi_r(\hat{Q}, \hat{\chi})}{\partial \hat{\chi}} \quad (39)$$

$$Q = -\frac{1}{\alpha} \frac{\partial \pi_r(\hat{Q}, \hat{\chi})}{\partial \hat{Q}} \quad (40)$$

$$\hat{\chi} = \frac{Q + \sigma_y^2 + m_y^2}{(1 + \chi)^2} \quad (41)$$

$$\hat{Q} = \frac{1}{1 + \chi}. \quad (42)$$

Note that the functional form of the parameters $\hat{\chi}$ and \hat{Q} does not depend on the regularization, but the values of χ and Q are regularization-dependent. At the

extremum, the parameters Q and χ are related to the physical quantities by

$$Q = \frac{1}{M} \sum_{i=1}^N E_{\mathbf{y}, \mathbf{A}} [||\hat{\mathbf{x}}(\mathbf{y}, \mathbf{A})||_2^2] \quad (43)$$

$$\chi = \lim_{\beta \rightarrow \infty} \frac{\beta}{M} \sum_{i=1}^N E_{\mathbf{y}, \mathbf{A}} [\langle ||\mathbf{x}||_2^2 \rangle_\beta - ||\langle \mathbf{x} \rangle_\beta||_2^2], \quad (44)$$

and can be expressed using x_r^* as

$$\chi = \frac{1}{\alpha} \int Dz \frac{\partial x_r^*(z; \hat{Q}, \hat{\chi})}{\partial (\sqrt{\hat{\chi}} z)} \quad (45)$$

$$Q = \frac{1}{\alpha} \int Dz (x_r^*(z; \hat{Q}, \hat{\chi}))^2. \quad (46)$$

The extended free energy density with the external fields κ and ν is given by

$$f_{\kappa, \nu} = f - \left\{ \frac{\alpha(m_y^2 + \sigma_y^2)\kappa(\kappa - 2)}{2(1 + \chi)} + \frac{\alpha\nu(\nu - 2m_y)}{2(1 + \chi)} \right\} \chi. \quad (47)$$

To evaluate $f_{\kappa, \nu}$ for non-zero κ and ν , one has to solve the saddle point equation at non-zero κ and ν to determine the saddle point value of χ . However, since one would only need to evaluate derivatives of $f_{\kappa, \nu}$ at $\kappa = \nu = 0$ to obtain GDF, the saddle point value of χ that is to be used in such evaluations should remain the same as that obtained in the calculation of f . From (22)–(24), GDF is obtained as

$$\text{df} = \frac{\chi}{1 + \chi} = \frac{\chi}{\hat{Q}^{-1}}, \quad (48)$$

where χ and \hat{Q} satisfy the saddle point equations (39) and (42), respectively. This expression is also independent of the form of the regularization. The effective single-body problem (37) can be interpreted as a scalar estimation problem in which x is estimated on the basis of the prior (regularization) $\exp(-r(x; \eta))$ and the random observation h/\hat{Q} , which is assumed to be generated as $h/\hat{Q} = x + n$, where $n \sim \mathcal{N}(0, \hat{Q}^{-1})$ is the Gaussian observation noise. If one uses the observation itself in the single-body problem as an estimate of x , then it is an unbiased estimator of x and its variance is \hat{Q}^{-1} . However, the actual variance of the estimates can change according to the regularization. The variable χ is the rescaled variance of the system expressed as (44). Therefore, GDF (48) corresponds to the effective fraction of the non-zero components of $\hat{\mathbf{x}}$ (parameters), which is estimated by dividing the variance of the total system by that of one component when the observation is used as the estimate. The effective fraction of the non-zero components is measured under the assumption that the regularization does not change the variance of one component from \hat{Q}^{-1} . If this assumption is correct and the fluctuation of non-zero components is the unique source of the system's fluctuation, GDF is considered to be equal to the ratio of the number of non-zero components to the number of training samples.

The RS solution discussed thus far loses local stability under perturbations that break the symmetry between replicas in a certain parameter region. Known as the de

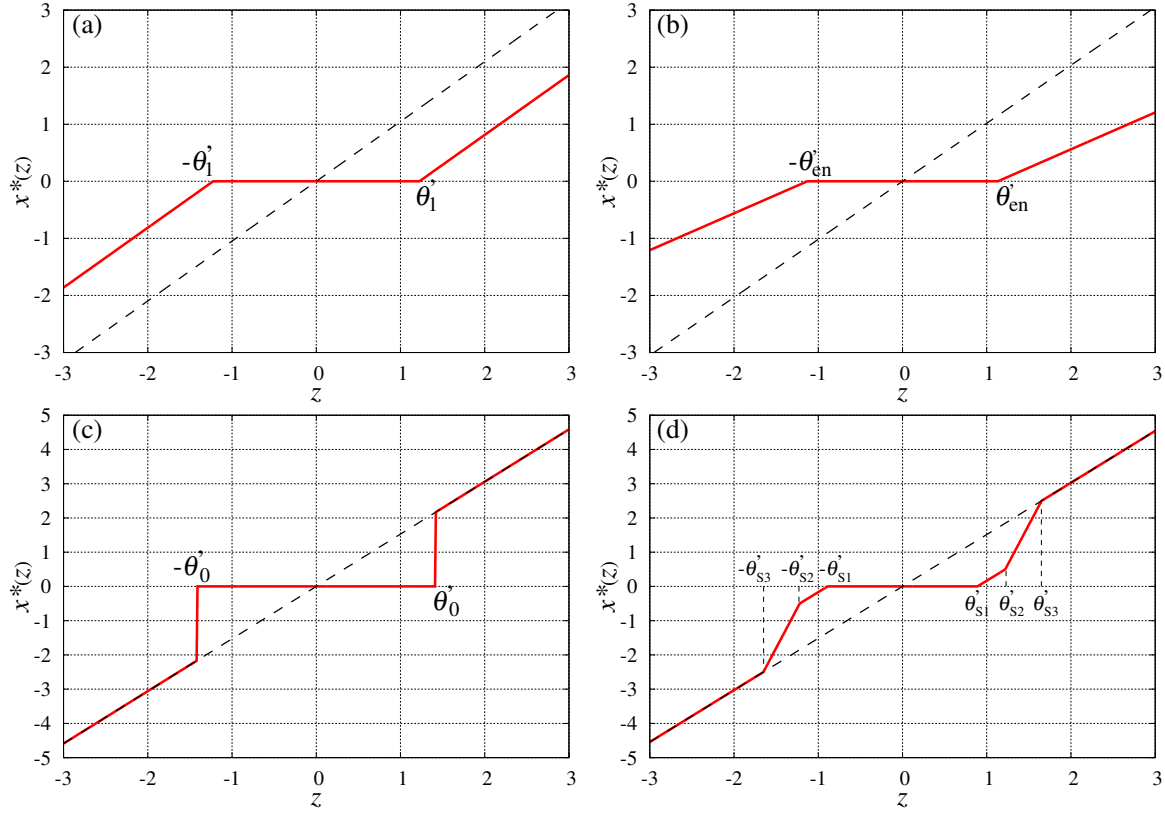


Figure 1. Behaviour of the maximizers of the single-body problem at $\alpha = 1$ for (a) ℓ_1 ($\eta = 1$), (b) elastic net ($\eta_1 = 1$, $\eta_2 = 0.5$), (c) ℓ_0 ($\eta = 1$), and (d) SCAD regularization ($\eta = 1$, $a = 5$, $\lambda = 0.5$). The dashed diagonal lines of gradient 1 are the maximizers under no regularization, and the threshold θ' denotes $\sqrt{2}\theta$.

Almeida–Thouless (AT) instability [35], this phenomenon appears when

$$\frac{1}{\alpha(1+\chi)^2} \int Dz \left\{ \frac{\partial x_r^*(z; \hat{Q}, \hat{\chi})}{\partial(\sqrt{\hat{\chi}}z)} \right\}^2 > 1. \quad (49)$$

In general, when AT instability appears, we have to construct the full-step replica symmetry breaking (RSB) solution for an exact evaluation. However, the RS solution remains meaningful as an approximation [32, 33].

5. Applications to several sparse regularizations

As shown in the previous section, some regularization-dependency appears in the effective single-body problem (37). We now apply the analytical method to ℓ_1 , elastic net, ℓ_0 , and SCAD regularization. The ratio of the number of non-zero components to the number of training samples is denoted by $\delta = \hat{\rho}/\alpha$, and we focus on the physical region $\delta \leq 1$, where the number of unknown variables is smaller than the number of known variables.

5.1. ℓ_1 regularization

In the ℓ_1 regularization $r(\mathbf{x}; \eta) = \eta \|\mathbf{x}\|_1 = \eta \sum_i |x_i|$, the maximizer of the single-body problem (37) is given by

$$x_{\ell_1}^*(z; \hat{Q}, \hat{\chi}) = \begin{cases} (h^{\text{RS}}(z; \hat{\chi}) - \eta \text{sgn}(z))/\hat{Q} & (|h^{\text{RS}}(z; \hat{\chi})| > \eta) \\ 0 & (\text{otherwise}) \end{cases}, \quad (50)$$

where $\text{sgn}(z)$ denotes the sign of z and is 0 when $z = 0$. Figure 1 (a) shows the behaviour of $x_{\ell_1}^*$ at $\alpha = 1$ and $\eta = 1$. Setting $\theta_1 = \eta/\sqrt{2\hat{\chi}}$, the fraction of non-zero components is given by the probability that the solution of the RS single-body problem (50) is non-zero: $\hat{\rho} = \text{erfc}(\theta_1)$, where

$$\text{erfc}(a) = \frac{2}{\sqrt{\pi}} \int_a^\infty dz e^{-z^2}, \quad (51)$$

and

$$\pi_{\ell_1} = \frac{\hat{\chi}}{\hat{Q}} \left\{ (1 + 2\theta_1^2)\hat{\rho} - \frac{2\theta_1}{\sqrt{\pi}} e^{-\theta_1^2} \right\}. \quad (52)$$

The regularization-dependent saddle point equations are given by

$$\chi = \frac{\hat{\rho}}{\alpha \hat{Q}} \quad (53)$$

$$Q = \frac{\hat{\chi} \hat{\tau}_1}{\alpha \hat{Q}^2}, \quad (54)$$

where

$$\hat{\tau}_1 = (1 + 2\theta_1^2)\hat{\rho} - \frac{2\theta_1}{\sqrt{\pi}} e^{-\theta_1^2}. \quad (55)$$

From (41) and (42), the solutions of the saddle point equations (53) and (54) can be derived as

$$\chi = \frac{\hat{\rho}}{\alpha - \hat{\rho}} \quad (56)$$

$$Q = \frac{(m_y^2 + \sigma_y^2)\hat{\tau}_1}{\alpha - \hat{\tau}_1}. \quad (57)$$

Substituting the saddle point equations, the free energy density and the expectation of the regularization term are given by

$$f = \frac{\alpha \hat{\chi}}{2} + \alpha(\chi \hat{\chi} - Q \hat{Q}) \quad (58)$$

$$\bar{r} = \lim_{N \rightarrow \infty} \frac{\eta}{N} E_{\mathbf{y}, \mathbf{A}} [\|\hat{\mathbf{x}}(\mathbf{y}, \mathbf{A})\|_1] = \alpha(\chi \hat{\chi} - Q \hat{Q}), \quad (59)$$

respectively. Hence, the training error is given by

$$\overline{\text{err}}_{\text{train}} = \hat{\chi}. \quad (60)$$

AT instability appears when

$$\frac{\hat{\rho}}{\alpha} > 1, \quad (61)$$

which is outside the region of interest for physical parameters. Equation (56) leads to the following expression for the GDF:

$$\text{df} = \frac{\hat{\rho}}{\alpha} = \delta. \quad (62)$$

This expression is consistent with the result in [19], which verifies the validity of this RS analysis for the derivation of GDF.

5.2. Elastic net regularization

Elastic net regularization, given by

$$r(\mathbf{x}; \eta_1, \eta_2) = \eta_1 \|\mathbf{x}\|_1 + \frac{\eta_2}{2} \|\mathbf{x}\|_2^2, \quad (63)$$

was developed to encourage the grouping effect, which is not exhibited by ℓ_1 regularization, and to stabilize the ℓ_1 regularization path [36]. Here, the coefficient $1/2$ is introduced for mathematical convenience, and $\eta_2 = 0$ and $\eta_1 = 0$ correspond to ℓ_1 regularization and ℓ_2 regularization, respectively.

The solution of the effective single-body problem for elastic net regularization is given by

$$x_{\text{en}}^*(z; \hat{Q}, \hat{\chi}) = \begin{cases} (h^{\text{RS}}(z; \hat{\chi}) - \eta_1 \text{sgn}(z)) / (\hat{Q} + \eta_2) & (|h^{\text{RS}}(z; \hat{\chi})| > \eta_1) \\ 0 & (\text{otherwise}) \end{cases} \quad (64)$$

The behaviour of this solution is shown in Fig. 1 (b) for $\alpha = 1$, $\eta_1 = 1$, $\eta_2 = 0.5$, and

$$\pi_{\text{en}} = \frac{\hat{\chi}}{\hat{Q} + \eta_2} \left\{ (1 + 2\theta_{\text{en}}^2) \text{erfc}(\theta_{\text{en}}) - \frac{2\theta_{\text{en}}}{\sqrt{\pi}} e^{-\theta_{\text{en}}^2} \right\}, \quad (65)$$

where $\theta_{\text{en}} = \eta_1 / \sqrt{2\hat{\chi}}$. The fraction of non-zero components is given by $\hat{\rho} = \text{erfc}(\theta_{\text{en}})$, and the regularization-dependent saddle point equations are given by

$$Q = \frac{\hat{\chi}}{\alpha(\hat{Q} + \eta_2)^2} \left\{ (1 + 2\theta_{\text{en}}^2) \hat{\rho} - \frac{2\theta_{\text{en}}}{\sqrt{\pi}} e^{-\theta_{\text{en}}^2} \right\} \quad (66)$$

$$\chi = \frac{\hat{\rho}}{\alpha(\hat{Q} + \eta_2)}. \quad (67)$$

At the saddle point, the free energy density and the expectation of the regularization term can be simplified as

$$f = \frac{\alpha\hat{\chi}}{2} + \alpha(\chi\hat{\chi} - Q\hat{Q}) - \frac{\alpha\eta_2 Q}{2} \quad (68)$$

$$\begin{aligned} \bar{r} &= \lim_{N \rightarrow \infty} \frac{1}{N} E_{\mathbf{y}, \mathbf{A}} \left[\eta_1 \|\hat{\mathbf{x}}\|_1 + \frac{\eta_2}{2} \|\hat{\mathbf{x}}\|_2^2 \right] \\ &= \alpha \{ \hat{\chi} \chi - (\hat{Q} + \eta_2) Q \} + \frac{\alpha\eta_2 Q}{2}, \end{aligned} \quad (69)$$

respectively. Hence, the training error is given by

$$\overline{\text{err}}_{\text{train}} = \hat{\chi}. \quad (70)$$

AT instability arises when

$$\frac{\hat{\rho}}{\alpha} > \left(\frac{\hat{Q} + \eta_2}{\hat{Q}} \right)^2; \quad (71)$$

the right-hand side is always greater than 1 because $\eta_2 \geq 0$ and $\hat{Q} > 0$. Therefore, the RS solution is always stable under symmetry breaking perturbations in the physical parameter region $\alpha > \hat{\rho}$.

From (67), the GDF for elastic net regularization is given by

$$\text{df} = \frac{\hat{\rho}}{\alpha} - \chi\eta_2 = \frac{\delta\hat{Q}}{\hat{Q} + \eta_2}, \quad (72)$$

which reduces to the GDF for ℓ_1 regularization at $\eta_2 = 0$. An unbiased estimator of the GDF for one instance of \mathbf{A} is derived in [20] as

$$\hat{\text{df}}(\mathbf{A}) = \frac{1}{M} \text{Tr}(\mathbf{A}_{\mathcal{A}}(\mathbf{A}_{\mathcal{A}}^T \mathbf{A}_{\mathcal{A}} + \eta_2 \mathbf{I}_{|\mathcal{A}|})^{-1} \mathbf{A}_{\mathcal{A}}^T), \quad (73)$$

where \mathcal{A} is the set of indices of non-zero components, and the columns $\{\mathbf{A}_i | i \in \mathcal{A}\}$ constitute the submatrix $\mathbf{A}_{\mathcal{A}}$. The number of the components of \mathcal{A} is denoted by $|\mathcal{A}|$. Our expression (72) for df corresponds to the typical value (or the expectation) of $\hat{\text{df}}(\mathbf{A})$ for a Gaussian random matrix \mathbf{A} . The physical implications suggested by the cavity method [32, 38], which is complementary to the replica method, supports the correspondence relationship between \hat{Q} in the replica method and the Gram matrix of \mathbf{A} . This correspondence indicates that our RS analysis is valid for the derivation of GDF under elastic net regularization.

As shown in (72), the GDF for elastic net regularization deviates from $\delta = \hat{\rho}/\alpha$. The ℓ_2 regularization term in elastic net regularization changes the variance of the non-zero components from \hat{Q}^{-1} to $(\hat{Q} + \eta_2)^{-1}$. Hence, the effective fraction of the non-zero components measured by χ/\hat{Q}^{-1} does not coincide with δ . By defining the rescaled estimates of the single-body problem as $x_{\text{en}}^{\text{res}} = (1 + \eta_2/\hat{Q})x_{\text{en}}^*$, the corresponding variance is reduced to $\chi^{\text{res}} = \hat{\rho}/(\alpha\hat{Q})$ from (45), and this gives $\text{df} = \delta$. This rescaling corresponds to that shown in [36], which was introduced to cancel out the shrinkage caused by ℓ_2 regularization and improve the prediction performance.

Taking the limit as $\eta_1 \rightarrow 0$, the GDF for ℓ_2 regularization can be obtained where the estimate is not sparse. The solution of the effective single-body problem is given by

$$x_{\ell_2}^*(z; \hat{Q}, \hat{\chi}) = \frac{h^{\text{RS}}(z; \hat{\chi})}{\hat{Q} + \eta_2}, \quad (74)$$

and the function π is given by

$$\pi_{\ell_2} = \frac{\hat{\chi}}{\hat{Q} + \eta_2}. \quad (75)$$

This expression leads to the following GDF:

$$\text{df} = \frac{\hat{Q}}{\hat{Q} + \eta_2}, \quad (76)$$

which corresponds to the limit as $\delta \rightarrow 1$ of the elastic net regularization. An unbiased estimator of the GDF for one instance of \mathbf{A} is proposed as [29]

$$\hat{\text{df}}(\mathbf{A}) = \frac{1}{M} \text{Tr} \mathbf{A}(\mathbf{A}^T \mathbf{A} + \eta_2 \mathbf{I}_N)^{-1} \mathbf{A}^T. \quad (77)$$

Equation (76) corresponds to the expectation of $\hat{\text{df}}(\mathbf{A})$ for a Gaussian random matrix \mathbf{A} .

5.3. ℓ_0 regularization

The ℓ_0 regularization is expressed by $r(\mathbf{x}; \eta) = \eta \|\mathbf{x}\|_0 = \eta \sum_i |x_i|_0$, which corresponds to the number of non-zero components in \mathbf{x} . The solution to the single-body problem for ℓ_0 regularization is given by

$$x_{\ell_0}^*(z; \hat{Q}, \hat{\chi}) = \begin{cases} h^{\text{RS}}(z; \hat{\chi})/\hat{Q} & (|h^{\text{RS}}(z; \hat{\chi})| > \sqrt{2\hat{\chi}}\theta_0) \\ 0 & (\text{otherwise}) \end{cases}, \quad (78)$$

where $\theta_0 = \sqrt{\eta\hat{Q}/\hat{\chi}}$, and by setting the fraction of non-zero components to $\hat{\rho} = \text{erfc}(\theta_0)$, we can derive

$$\pi_{\ell_0} = \frac{\hat{\chi}}{\hat{Q}} \left\{ \frac{2\theta_0}{\sqrt{\pi}} e^{-\theta_0^2} + (1 - 2\theta_0^2)\hat{\rho} \right\}. \quad (79)$$

Figure 1 (c) shows the z -dependence of the maximizer $x_{\ell_0}^*$ at $\alpha = 1$ and $\eta = 1$. The regularization-dependent saddle point equations (39)–(40) are given by

$$\chi = \frac{1}{\alpha\hat{Q}} \left\{ \frac{2\theta_0}{\sqrt{\pi}} e^{-\theta_0^2} + \hat{\rho} \right\} \quad (80)$$

$$Q = \frac{\hat{\chi}}{\alpha\hat{Q}^2} \left\{ \frac{2\theta_0}{\sqrt{\pi}} e^{-\theta_0^2} + \hat{\rho} \right\}, \quad (81)$$

and have two solutions: finite χ and Q and infinite χ and Q . We denote the finite and infinite solutions as $S_1 = \{\chi_1, Q_1\}$ and $S_2 = \{\chi_2 = \infty, Q_2 = \infty\}$, respectively. Using (41) and (42), the finite solution can be simplified as

$$\chi_1 = \frac{\hat{\rho} + \omega}{\alpha - (\hat{\rho} + \omega)} \quad (82)$$

$$Q_1 = (m_y^2 + \sigma_y^2)\chi_1, \quad (83)$$

where

$$\omega = \int Dz |z| \delta(|z| - \sqrt{2}\theta_0) = \frac{2\theta_0}{\sqrt{\pi}} e^{-\theta_0^2}. \quad (84)$$

By definition, χ_1 and Q_1 should be positive, and so (82)–(83) are only valid when $\alpha > \hat{\rho} + \omega$. According to a local stability analysis of (80) around $1/\chi = 0$, solution S_2 is a locally stable solution of the RS saddle point equation when $\alpha < \hat{\rho} + \omega$, where as it is unstable when $\alpha > \hat{\rho} + \omega$. Therefore, the stable solution of the RS saddle point equation changes from S_1 to S_2 at $\alpha = \hat{\rho} + \omega$. Note that the stability discussed here refers to the RS solution, and does not relate to AT instability.

The free energy density is simplified by substituting the saddle point equations as

$$f = \frac{\alpha\hat{\chi}}{2} + \eta\hat{\rho}. \quad (85)$$

The second term of (85) corresponds to the expectation of the regularization term, and so the training error can be derived as

$$\overline{\text{err}}_{\text{train}} = \hat{\chi}. \quad (86)$$

The GDF is given by

$$\text{df} = \begin{cases} \delta + \frac{\omega}{\alpha} & \text{for solution } S_1 \\ 1 & \text{for solution } S_2 \end{cases}. \quad (87)$$

The term ω , given by (84), in the GDF originates from the discontinuity of the single-body problem at the threshold $\sqrt{2}\theta_0$, as shown in figure 1 (c). In addition to the fluctuation generated by the non-zero components, this discontinuity induces fluctuations in the system and increases the GDF from δ .

Under ℓ_0 regularization, AT instability always appears, but the estimated GDF under the RS assumption can be regarded as an approximation of the true value of the GDF, as shown in Sec. 8. Our calculations based on the one-step RSB assumption indicate that the form of the GDF, as the fraction of non-zero components plus the discontinuity term, is unchanged, although the values of these two terms does change (unreported).

5.4. SCAD regularization

SCAD regularization is a non-convex sparse regularization in which the estimator has the desirable properties of being unbiased, sparse, and continuous [10]. Mathematically, the SCAD estimator is asymptotically equivalent to the oracle estimator [10, 11]. SCAD regularization is given by

$$r(x; \eta) = \begin{cases} \eta\lambda|x| & (|x| \leq \lambda) \\ -\eta\left\{\frac{x^2 - 2a\lambda|x| + \lambda^2}{2(a-1)}\right\} & (\lambda < |x| \leq a\lambda) \\ \frac{\eta(a+1)\lambda^2}{2} & |x| > a\lambda \end{cases}, \quad (88)$$

where λ and a are parameters that control the form of the regularization. The maximizer of the single-body problem for SCAD regularization is given by

$$x_S^*(z; \hat{Q}, \hat{\chi}) = \begin{cases} \frac{h^{\text{RS}}(z; \hat{\chi}) - \lambda\eta\text{sgn}(z)}{\hat{Q}} & (\lambda\eta < |h^{\text{RS}}(z; \hat{\chi})| \leq \lambda(\hat{Q} + \eta)) \\ \frac{h^{\text{RS}}(z; \hat{\chi})(a-1) - a\lambda\eta\text{sgn}(z)}{\hat{Q}(a-1) - \eta} & (\lambda(\hat{Q} + \eta) < |h^{\text{RS}}(z; \hat{\chi})| \leq a\lambda\hat{Q}) \\ \frac{h^{\text{RS}}(z; \hat{\chi})}{\hat{Q}} & (|h^{\text{RS}}(z; \hat{\chi})| > a\lambda\hat{Q}) \\ 0 & (\text{otherwise}) \end{cases}. \quad (89)$$

Figure 1 (d) shows an example of the behaviour of the maximizer x_S^* at $a = 5$, $\lambda = 0.1$, and $\eta = 1$, where three thresholds are given by $\theta_{S1} = \lambda\eta/\sqrt{2\hat{\chi}}$, $\theta_{S2} = \lambda(\hat{Q} + \eta)/\sqrt{2\hat{\chi}}$, and $\theta_{S3} = a\lambda\hat{Q}/\sqrt{2\hat{\chi}}$. The threshold θ_{S1} gives the fraction of non-zero components as $\hat{\rho} = \text{erfc}(\theta_{S1})$. Between the thresholds θ_{S1} and θ_{S2} , and beyond the third threshold θ_{S3} , the estimate x_S^* behaves like the ℓ_1 and ℓ_0 estimates, respectively. Between θ_{S2} and θ_{S3} , the estimate transits linearly between the ℓ_1 and ℓ_0 estimates.

The function π for SCAD regularization is derived as

$$\pi_S = \pi_1 + \pi_2 + \pi_3 + \frac{\eta\lambda^2\pi_4}{a-1} - \eta(a+1)\lambda^2\text{erfc}(\theta_{S3}), \quad (90)$$

where

$$\pi_1 = \frac{\hat{\chi}}{\hat{Q}} \left[-\frac{2\theta_{S1}}{\sqrt{\pi}} \left(e^{-\theta_{S1}^2} + \left(\frac{\hat{Q}-\eta}{\eta} \right) e^{-\theta_{S2}^2} \right) + (1+2\theta_{S1}^2) \{ \hat{\rho} - \text{erfc}(\theta_{S2}) \} \right] \quad (91)$$

$$\begin{aligned} \pi_2 = & \frac{\hat{\chi}}{\hat{Q} - \frac{\eta}{a-1}} \left[\frac{2}{\sqrt{\pi}} \left\{ \left(\theta_{S2} - \frac{2\theta_{S3}\eta}{\hat{Q}(a-1)} \right) e^{-\theta_{S2}^2} - \left(1 - \frac{2\eta}{\hat{Q}(a-1)} \right) \theta_{S3} e^{-\theta_{S3}^2} \right\} \right. \\ & \left. + \left\{ 1 + 2 \left(\frac{\eta\theta_{S3}}{\hat{Q}(a-1)} \right)^2 \right\} \pi_4 \right] \end{aligned} \quad (92)$$

$$\pi_3 = \frac{\hat{\chi}}{\hat{Q}} \left[\frac{2\theta_{S3}}{\sqrt{\pi}} e^{-\theta_{S3}^2} + \text{erfc}(\theta_{S3}) \right] \quad (93)$$

$$\pi_4 = \text{erfc}(\theta_{S2}) - \text{erfc}(\theta_{S3}). \quad (94)$$

The regularization-dependent saddle point equations are given by

$$Q = \frac{1}{\alpha} \left\{ \frac{\pi_1}{\hat{Q}} + \frac{\pi_2}{\hat{Q} - \frac{\eta}{a-1}} + \frac{\pi_3}{\hat{Q}} \right\} \quad (95)$$

$$\chi = \frac{1}{\alpha\hat{Q}} \left[\hat{\rho} + \frac{\frac{\eta}{a-1}}{\hat{Q} - \frac{\eta}{a-1}} \pi_4 \right], \quad (96)$$

and the expectation of the regularization term is given by

$$\bar{r} = \alpha\chi\hat{\chi} - \pi_1 - \left\{ 1 + \frac{\frac{\eta}{a-1}}{2(\hat{Q} - \frac{\eta}{a-1})} \right\} \pi_2 - \pi_3 - \frac{\eta\lambda^2}{2(a-1)} \pi_4 + \frac{\eta(a+1)\lambda^2}{2} \text{erfc}(\theta_{S3}). \quad (97)$$

Substituting these equations into the free energy density, we get

$$\overline{\text{err}}_{\text{train}} = \hat{\chi}. \quad (98)$$

The AT instability condition is given by

$$\frac{1}{\alpha(1+\chi)^2} \left[\frac{\hat{\rho}}{\hat{Q}^2} + \left\{ \left(\hat{Q} - \frac{\eta}{a-1} \right)^{-2} - \frac{1}{\hat{Q}^2} \right\} \pi_4 \right] > 1, \quad (99)$$

which reduces to that for ℓ_1 regularization as $a \rightarrow \infty$.

There are three solutions of $\{Q, \chi\}$: $S_1 = \{Q = Q_1 < \infty, \chi = \chi_1 < \infty\}$, $S_2 = \{Q = Q_2 < \infty, \chi = \infty\}$, and $S_3 = \{Q = \infty, \chi = \infty\}$. For sufficiently large a , the finite solution S_1 is a locally stable solution of the RS saddle point equation when

$$\alpha > \hat{\rho} + \frac{\eta/\{\hat{Q}(a-1)\}}{1 - \eta/\{\hat{Q}(a-1)\}} \pi_4. \quad (100)$$

Beyond the range of (100), the stable RS solution is replaced by S_2 . For sufficiently small η , the stable RS solution can switch from S_2 to S_3 depending on the SCAD parameter,

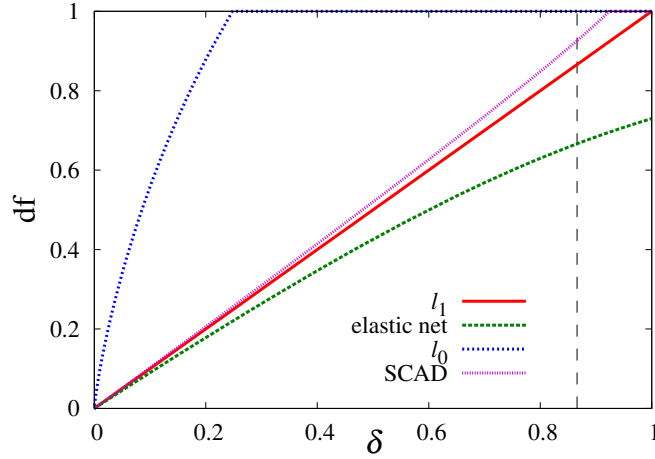


Figure 2. δ -dependence of GDF for ℓ_1 , elastic net, ℓ_0 , and SCAD regularization at $\alpha = 0.5$, $m_y = 0$, and $\sigma_y^2 = 1$. The parameters for elastic net and SCAD regularization are $\eta_2 = 0.1$, $a = 8$, and $\lambda = 1$, and the vertical dashed line indicates the appearance of the AT instability for SCAD regularization, $\delta = 0.866$. The ℓ_1 result corresponds to the line $df = \delta$.

but this is not important in estimating the GDF, because both solutions give the same GDF value. The GDF for SCAD regularization can be summarized as

$$df = \begin{cases} \delta + \frac{\frac{\eta}{a-1}}{\alpha(\hat{Q} - \frac{\eta}{a-1})} \pi_4 & \text{for solution S}_1 \\ 1 & \text{for solutions S}_2 \text{ and S}_3 \end{cases}. \quad (101)$$

As $a \rightarrow \infty$, we get $\pi_4 \rightarrow 0$ and solution S_1 is always a stable RS solution, satisfying (100); hence, the GDF reduces to that for ℓ_1 regularization. The second term of the GDF for solution S_1 arises from the weight between the thresholds θ_{S_2} and θ_{S_3} . The manner of assigning the non-zero components to this transient region between the ℓ_1 and ℓ_0 estimates increases the fluctuation in the system, and the GDF does not coincide with δ .

We note the pathology of solution S_3 under the RS assumption. As shown in the solution to the single-body problem (89) (Fig. 1 (d)), the magnitude relation $\theta_{S_1} \leq \theta_{S_2} \leq \theta_{S_3}$ should hold. However, the $Q, \chi \rightarrow \infty$ solution leads to $\theta_{S_3} \rightarrow 0$ with finite $\theta_{S_1} = \theta_{S_2}$. Solution S_3 appears in the region where AT instability appears, and so this non-physical phenomenon is considered to be caused by an inappropriate RS assumption. Hence, we must construct the RSB solution to correctly describe the GDF corresponding to solution S_3 .

6. Parameter dependence of GDF and prediction error

Figure 2 illustrates the δ -dependence of the GDF for ℓ_1 , the elastic net with $\eta_2 = 0.1$, ℓ_0 , and SCAD regularization with $a = 8$ and $\lambda = 1$ at $\alpha = 0.5$, $m_y = 0$, and $\sigma_y^2 = 1$. At each point of δ , the regularization parameters η for ℓ_1 , ℓ_0 and SCAD regularization and η_2

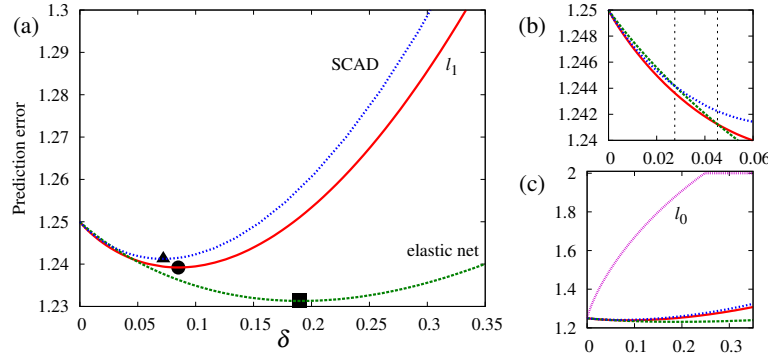


Figure 3. (a) δ -dependence of the prediction error $\overline{\text{err}}_{\text{pre}}$ for ℓ_1 , elastic net, and SCAD regularization at $\alpha = 0.5$, $m_y = 0.5$, and $\sigma_y^2 = 1$. The parameters for elastic net and SCAD regularization are $\eta_2 = 0.1$, $a = 8$, and $\lambda = 1$. (b) Region where the magnitude relationship between each regularization changes. (c) Prediction error for ℓ_0 regularization.

for elastic net regularization are controlled such that $\delta = \hat{\rho}/\alpha$. Under ℓ_1 regularization, the GDF is always equal to δ , as shown in (62). In elastic net regularization, the GDF is less than δ as the ℓ_2 parameter η_2 increases. For ℓ_0 regularization, the RS solution S_1 is unstable at $\delta > 0.248$ in this parameter region, and is replaced by solution S_2 , which gives $\text{df} = 1$. In SCAD regularization, the solution S_1 loses local stability within the RS assumption at $\delta > 0.924$, and AT instability appears before the RS solution S_1 becomes unstable at $\delta > 0.866$ (denoted by the dashed vertical line in figure 2.)

Figure 3 shows the prediction error (26) for the same parameter region as figure 2. At $\sigma_y^2 = 1$, the prediction error is equivalent to the expectation of AIC. In the entire range of δ shown in figure 3 (a), the RS solutions for ℓ_1 , elastic net, and SCAD regularization are stable under symmetry breaking perturbations. Thus, we can identify the value of δ that minimizes the prediction error for each regularization. In this case, the models with $\delta = 0.085$ (denoted by \bullet), $\delta = 0.170$ (\blacksquare), and $\delta = 0.072$ (\blacktriangle) are selected for ℓ_1 , elastic net, and SCAD regularization, respectively. In the current problem setting, sparse estimation with SCAD regularization minimizes the prediction error within the RS region when the mean of the data is sufficiently small. To identify the appropriate model using RS analysis, it is useful to standardize the data. As shown in figure 3 (b), the magnitude of the prediction errors at $\delta < \delta_1 = 0.028$, $\delta_1 < \delta < \delta_2 = 0.045$, and $\delta > \delta_2$ runs in descending order as elastic net $>$ SCAD $>$ ℓ_1 , SCAD $>$ elastic net $>$ ℓ_1 , and SCAD $>$ ℓ_1 $>$ elastic net, respectively. The estimates $\hat{\mathbf{x}}$ have different supports depending on the regularization, even when the regularization parameters are controlled to give a certain value of δ . A comparison of the prediction errors within the framework of RS analysis guides the choice of regularization for each value of δ .

The prediction error for ℓ_0 regularization under the RS assumption is shown in figure 3 (c) alongside those for other regularization types. The RS prediction error is minimized at $\delta = 0$. This indicates that the appropriate model under RS analysis

has a non-zero component of $O(1)$. Our analysis assumes that the number of non-zero components is $O(N)$; hence, the derived model selection criterion cannot identify the appropriate model in the current problem setting for ℓ_0 regularization.

7. Numerical calculation of GDF using belief propagation algorithm

7.1. Belief propagation algorithm for sparse regularization

The correspondence between replica analysis and the belief propagation (BP) algorithm suggests that the typical properties of BP fixed points at the large-system-size limit can be described by the RS saddle point [32, 38]. Thus, we may expect that the numerically obtained GDF will be consistent with the RS analysis at finite system sizes using the BP algorithm. For the ordinary least squares with a regularization that can be written as (12), a tentative estimate of the i -th component at step t , denoted by $\hat{x}_i^{(t)}$, is given by the solution to the single-body problem (37) with the substitutions $\hat{Q} \rightarrow \hat{Q}_i^{(t)}$ and $h^{\text{RS}}(z; \hat{\chi}) \rightarrow h_i^{(t)}$ [25, 26, 27], where

$$h_i^{(t)} = \hat{x}_i^{(t-1)} \sum_{\mu=1}^M \frac{A_{\mu i}^2}{1 + \sigma_{\mu}^{(t-1)^2}} + \sum_{\mu=1}^M A_{\mu i} R_{\mu}^{(t-1)} \quad (102)$$

$$\hat{Q}_i^{(t)} = \sum_{\mu=1}^M \frac{A_{\mu i}^2}{1 + \sigma_{\mu}^{(t-1)^2}}, \quad (103)$$

and setting $\hat{\mathbf{y}}^{(t)} = \mathbf{A}\hat{\mathbf{x}}^{(t)}$,

$$R_{\mu}^{(t)} = \frac{y_{\mu} - \hat{y}_{\mu}^{(t)}}{1 + \sigma_{\mu}^{(t)^2}} \quad (104)$$

$$\sigma_{\mu}^{(t)^2} = \frac{1}{\alpha} \sum_i A_{\mu i}^2 \chi_i^{(t)}. \quad (105)$$

The variable $\chi_i^{(t)}$ represents the variance of $x_i^{(t)}$, and its determination rule depends on the regularization. For ℓ_1 and elastic net regularization, the variable is given by

$$\chi_i^{(t)} = \begin{cases} \frac{1}{\hat{Q}_i^{(t)}} & \text{for } |h_i^{(t)}| > \eta \\ 0 & \text{otherwise} \end{cases} \quad (106)$$

and

$$\chi_i^{(t)} = \begin{cases} \frac{1}{\hat{Q}_i^{(t)} + \eta_2} & \text{for } |h_i^{(t)}| > \eta_1 \\ 0 & \text{otherwise} \end{cases}, \quad (107)$$

respectively. For these regularizations, the GDF at the BP fixed point converges to that given by RS analysis as the system size increases. However, for these regularizations, the GDF can be calculated using least angle regression (LARS) [37], which has a lower computational cost than the BP algorithm. Hence, there is no need to introduce the

BP algorithm. In the case of ℓ_0 regularization, the variable χ_i is given by

$$\chi_i^{(t)} = \begin{cases} \frac{1}{\hat{Q}_i^{(t)}} & \text{for } |h_i^{(t)}| > \sqrt{2\eta\hat{Q}_i^{(t)}} \\ 0 & \text{otherwise} \end{cases}. \quad (108)$$

Unfortunately, AT instability appears across the whole parameter region for ℓ_0 regularization, and the BP algorithm does not converge.

For SCAD regularization, no numerical method for the precise evaluation of GDF has been proposed. As shown in the previous section, SCAD regularization gives a parameter region where the RS solution is stable. Therefore, the BP algorithm is useful as a method of numerically calculating the GDF for SCAD regularization. The variable $\chi_i^{(t)}$ for SCAD regularization is given by

$$\chi_i^{(t)} = \begin{cases} \frac{1}{\hat{Q}_i^{(t)}} & \text{for } \lambda\eta < |h_i^{(t)}| \leq \lambda(\hat{Q}_i^{(t)} + \eta) \text{ and } |h_i^{(t)}| > a\lambda\hat{Q}_i^{(t)} \\ \frac{1}{\hat{Q}_i^{(t)} - \frac{\eta}{a-1}} & \text{for } \lambda(\hat{Q}_i^{(t)} + \eta) < |h_i^{(t)}| \leq a\lambda\hat{Q}_i^{(t)} \\ 0 & \text{otherwise} \end{cases} \quad (109)$$

After updating the estimates $\hat{\mathbf{x}}^{(t)}$, we can numerically evaluate the value of GDF using (8) with the data estimates $\hat{\mathbf{y}}^{(t)}$. To ensure convergence, appropriate damping is required at each update.

As for the replica analysis, we apply the BP algorithm for the case of Gaussian random data \mathbf{y} and predictors \mathbf{A} . Figure 4 shows the numerically calculated GDF given by the BP algorithm at $N = 200$, $m_y = 0$, and $\sigma_y^2 = 1$. The BP algorithm is updated until $|\hat{x}_i^{(t)} - \hat{x}_i^{(t-1)}| < 10^{-10}$ for each component, and the result is averaged over 100 realizations of $\{\mathbf{y}, \mathbf{A}\}$. The solid and dashed lines represent the analytical results given by the replica method for the RS and RSB regions, respectively. In the RS regime, the numerically calculated GDF from the BP algorithm coincides with that evaluated by the replica method.

7.2. Perspective for other predictors

The RS analysis discussed so far has been applied to Gaussian i.i.d. random predictors. Its extension to other predictors is not straightforward. To check the generality of the GDF being given by the effective fraction of non-zero components (χ/\hat{Q}^{-1}) at the RS saddle point for other predictor matrices, we resort to the BP algorithm. The typical properties of $\chi_i^{(t)}$ and $\hat{Q}_i^{(t)}$ at the BP fixed point denoted by χ_i^* and \hat{Q}_i^* are described in the replica method by χ and \hat{Q} of the RS saddle point at the large-system-size limit. Therefore, it is reasonable to define the effective fraction of non-zero components at the BP fixed point as

$$\delta_{\text{eff}}^{\text{BP}} = \frac{\overline{\frac{1}{N} \sum_i \chi_i^*(\mathbf{y}, \mathbf{A})}}{\left(\overline{\frac{1}{N} \sum_i \hat{Q}_i^*(\mathbf{y}, \mathbf{A})}\right)^{-1}}, \quad (110)$$

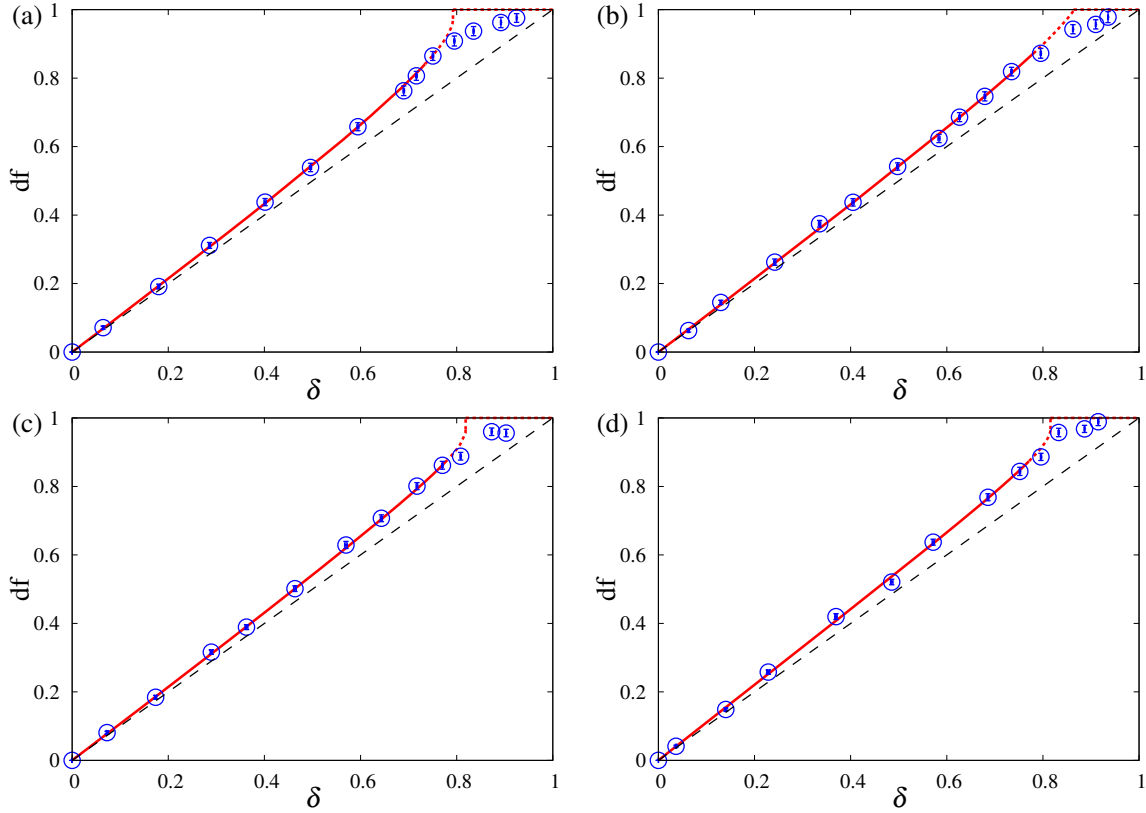


Figure 4. Comparison between BP algorithm and RS analysis for SCAD regularization at $N = 200$, $m_y = 0$, and $\sigma_y^2 = 1$ for (a) $a = 5$, $\lambda = 1$ for $M = 100$ ($\alpha = 0.5$), (b) $a = 8$, $\lambda = 0.8$ for $M = 100$ ($\alpha = 0.5$), (c) $a = 6$, $\lambda = 0.9$ for $M = 160$ ($\alpha = 0.8$), and (d) $a = 8$, $\lambda = 0.7$ for $M = 160$ ($\alpha = 0.8$). The BP results, averaged over 100 realizations of $\{\mathbf{y}, \mathbf{A}\}$, are denoted by circles. The theoretical estimation of GDF by RS analysis is denoted by solid and dashed lines for the RS and RSB regions, respectively. To provide a visual guide, the dashed line has a gradient of 1.

where the overline represents the average over \mathbf{y} and \mathbf{A} . If $\delta_{\text{eff}}^{\text{BP}}$ and the GDF from (8) coincide at the BP fixed point, it is considered that the correspondence between GDF and the effective fraction of non-zero components holds at the RS saddle point. For ℓ_1 , elastic net, and SCAD regularization, we examine the behaviour of GDF and $\delta_{\text{eff}}^{\text{BP}}$ under two predictors [36] in a parameter region where the BP algorithm converges.

Example 1: Gaussian predictors with pairwise correlation. The correlation between predictors \mathbf{A}_i and \mathbf{A}_j is set to be $c^{|i-j|}$, and the predictors are normalized such that $\|\mathbf{A}_i\|_2^2 = 1$.

Example 2: The predictors are generated as

$$\mathbf{A}_i = \begin{cases} \mathbf{Z}_1 + \boldsymbol{\epsilon}_i & \text{for } i = 1, \dots, T \\ \mathbf{Z}_2 + \boldsymbol{\epsilon}_i & \text{for } i = K + 1, \dots, 2T \\ \mathbf{Z}_3 + \boldsymbol{\epsilon}_i & \text{for } i = 2K + 1, \dots, 3T \\ \mathbf{a}_i & \text{for } i = 3T + 1, \dots, N \end{cases}, \quad (111)$$

where the components of the M -dimensional vectors $\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3$, $\{\mathbf{a}_i\}$, and $\{\boldsymbol{\epsilon}_i\}$

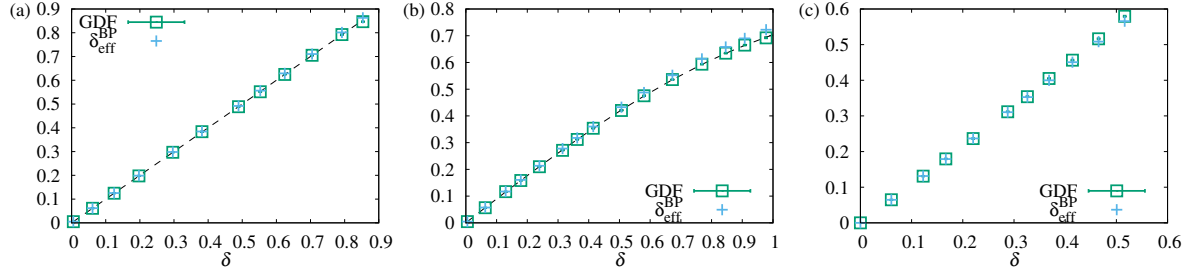


Figure 5. δ -dependence of GDF and $\delta_{\text{eff}}^{\text{BP}}$ at BP fixed point of $N = 1000$, $M = 500$ ($\alpha = 0.5$) for (a) ℓ_1 , (b) elastic net of $\eta_2 = 0.1$, and (c) SCAD regularization of $a = 0.5$ and $\lambda = 1$ under the predictor matrix of example 1 with $c = 0.5$. We used 1000 samples of predictor matrices to calculate the GDF and $\delta_{\text{eff}}^{\text{BP}}$ at the BP fixed point. The δ -region where the BP algorithm converges within 10^5 steps is shown. The dashed lines in (a) and (b) denote the results reported in [19] and [20].

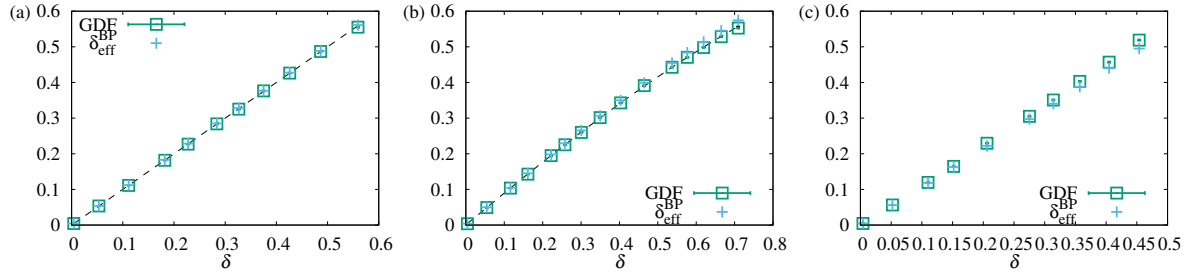


Figure 6. δ -dependence of GDF and $\delta_{\text{eff}}^{\text{BP}}$ at BP fixed point of $N = 1000$, $M = 500$ ($\alpha = 0.5$) for (a) ℓ_1 , (b) elastic net of $\eta_2 = 0.1$, and (c) SCAD regularization of $a = 0.5$ and $\lambda = 1$ under the predictor matrix of example 2 with $T = 125$. We used 1000 samples of predictor matrices to calculate the GDF and $\delta_{\text{eff}}^{\text{BP}}$ at the BP fixed point. The δ -region where the BP algorithm converges within 10^5 steps is shown. The dashed lines in (a) and (b) denote the results reported in [19] and [20].

are i.i.d. Gaussian random variables with mean zero and variance 1, and T is a parameter that takes an integer value smaller than $(N - 1)/3$. The predictors are normalized such that $\|\mathbf{A}_i\|_2^2 = 1$.

Figures 5 and 6 show the δ -dependence of GDF and $\delta_{\text{eff}}^{\text{BP}}$ at the BP fixed point for ℓ_1 , elastic net, and SCAD regularization at $N = 1000$ and $\alpha = 0.5$ ($M = 500$). The values of each point have been averaged over 1000 samples of $\{\mathbf{y}, \mathbf{A}\}$. Under ℓ_1 and elastic net regularization, the GDF value calculated as the expectation of the unbiased estimator derived in [19, 20] is shown as a dashed line. In both examples, the correspondence between GDF and $\delta_{\text{eff}}^{\text{BP}}$ holds for each regularization, although a small discrepancy appears due to finite-size effects at large δ . Furthermore, the values of $\delta_{\text{eff}}^{\text{BP}}$ and GDF at the BP fixed point are consistent with those of previous studies for ℓ_1 and elastic net regularization. The parameters c and T in these examples do not influence the results, although they do affect the convergence of the BP algorithm. These results imply that the correspondence between GDF and the effective fraction of

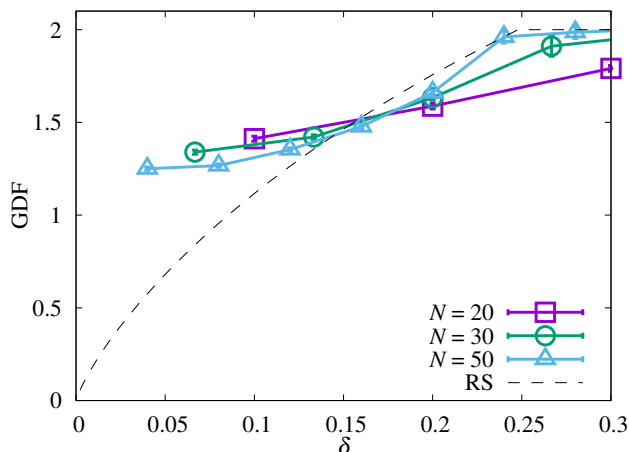


Figure 7. Comparison between exact value of GDF at finite system sizes and GDF under RS analysis.

non-zero components holds outside of Gaussian i.i.d. predictors. For both examples, the convergence of the BP algorithm is worse than with the Gaussian i.i.d. predictors, particularly at large δ . Thus, the algorithm must be improved to enable a discussion of the large- δ region and application to other predictor matrices.

In the case of ℓ_0 regularization, the BP algorithm does not converge. Thus, we cannot confirm the generality of the result using the properties of BP fixed points. The replica analysis for non-Gaussian i.i.d. predictor matrices is a necessary step towards verifying the generality of the result for ℓ_0 regularization. Although the range of applicable predictor matrices for replica analysis is narrower than that for the BP algorithm, the analysis of rotationally invariant predictor matrices offers a promising means towards demonstrating the generality [21].

8. RS solution approximates the GDF for ℓ_0 regularization

For ℓ_0 regularization, the RS solution is unstable in the whole parameter region, but it is known that this solution generally approximates the true solution. One can numerically obtain the exact solution of (12) for ℓ_0 regularization by an exhaustive search, and calculate the exact value of GDF at small system sizes. Comparing the GDF under RS analysis with its exact value, we can evaluate the approximation performance of the RS solution.

Figure 7 compares the GDF approximated by the RS solution with its exact value for $N = 20, 30$, and 50 as calculated by 1000 samples of $\{\mathbf{y}, \mathbf{A}\}$. As N increases, the exact GDF approaches the RS solution, although intense finite-size effects are observed in the small- δ region. For a comparison at larger system sizes, we must develop a computationally feasible algorithm for obtaining precise solutions of (12) for ℓ_0 regularization, but this is beyond the scope of the present paper.

9. Summary and Conclusion

We have derived the GDF using a method based on statistical physics. Within the range of the RS assumption, the GDF is represented as χ/\hat{Q}^{-1} , where χ and \hat{Q}^{-1} correspond to the rescaled variance around estimates and the variance of estimates when the regularization term is omitted, respectively. This expression does not depend on the type of regularization, and indicates that GDF can be regarded as the effective fraction of non-zero components.

We applied our method for the derivation of GDF to ℓ_1 , elastic net, ℓ_0 , and SCAD regularization. Our RS analysis was stable for ℓ_1 and elastic net regularization in the entire physical parameter region, and the GDFs for these regularizations were consistent with previous results. This correspondence supports the validity of our RS analysis. The model selection criterion of prediction error was derived by combining the GDF with the training error. Theoretical predictions in the RS phase were then algorithmically achieved using the belief propagation method.

It has been implied that the equivalence between GDF and the ratio of the number of non-zero components to the number of samples, δ , only holds for ℓ_1 regularization [19]. Our representation of GDF as the effective fraction of non-zero components clarifies the origin of the additional component of the GDF from the fraction of non-zero components.

- In ℓ_1 regularization, the GDF is given by δ because there is no factor that induces fluctuations other than the non-zero components.
- Elastic net regularization changes the variance of the components, and so the GDF does not coincide with δ . However, as with ℓ_1 regularization, the non-zero components are the unique source of fluctuations, and so the correspondence between GDF and δ can be recovered by appropriately rescaling the estimates.
- In ℓ_0 regularization, the discontinuity of the estimates leads to additional fluctuations besides those caused by the non-zero components. Hence, the GDF is greater than δ .
- In SCAD regularization, the assignment of non-zero components to the transient region between ℓ_1 -type estimates and ℓ_0 -type estimates induces additional components in the GDF.

For regularizations with AT instabilities in certain parameter regions (e.g. ℓ_0 , SCAD, and other non-convex regularizations), it is generally necessary to construct the full-step RSB solution. In the case of SCAD regularization, model selection based on the prediction error under RS analysis can be achieved in the current problem setting. Even when the RS solution is unstable, the prediction error gives a meaningful approximation of the true value.

Further development of our method for the general function of prediction error [39] and real data will be useful for practical applications. The BP algorithm discussed here can numerically calculate the GDF and model selection criterion for practical settings at reasonable computational cost.

Acknowledgments

The author would like to thank Yukito Iba, Yoshiyuki Kabashima, and Yoshiyuki Ninomiya for insightful discussions and comments. This work was supported by JSPS KAKENHI No.25120013, 26880028 and 16K16131.

References

- [1] Akaike H 1973 *Second International Symposium on Information Theory* (Petrov B N and Csaki F, eds.) 267 Académiai Kiadó, Budapest.
- [2] Tibshirani R 1996 *J. Roy. Statist. Soc. Ser. B* **58** 267
- [3] Candès E J and Tao T 2005 *IEEE Trans. Inform. Theory* **51** 4203
- [4] Donoho D 2006 *IEEE Trans. Inform. Theory* **52** 1289
- [5] Girolami M 2001 *Neural Comput.* **13** 2517
- [6] Zhu J, Rosset S, Hastie T and Tibshirani R 2004 *Adv. Neural. Inf. Process* **16** 49
- [7] Foucart S and Lai M-J 2009 *Appl. Comput. Harmon. Anal.* **26** 395
- [8] Wang M, Xu W and Tang A 2011 *IEEE Trans. Inform. Theory* **57** 7255
- [9] Xu Z, Chang X, Xu F and Zhang H 2012 *IEEE Trans. Neural Networks and Learning Systems* **23** 1013
- [10] Fan J and Li R 2001 *J. Amer. Statist. Assoc.* **96** 1348
- [11] Fan J and Peng H 2004 *Annal. Stat.* **32** 928
- [12] Huang J, Ma S and Zhang C-H 2008 *Statistica Sinica* **18** 1603
- [13] Stone M 1974 *Biometrika* **61** 509
- [14] Zhang Y, Li R and Tsai C-L 2010 *J. Amer. Statist. Assoc.* **105** 312
- [15] Obuchi T and Kabashima Y arXiv:1601.00881
- [16] Ye J 1998 *J. Amer. Statist. Assoc.* **93** 120
- [17] Mallows C 1973 *Technometrics* **15** 661
- [18] Efron B 2004 *J. Amer. Statist. Assoc.* **99** 619
- [19] Zou H, Hastie T and Tibshirani R 2007 *Annal. Stat.* **35** 2173
- [20] Zou H 2005 *Ph.D. dissertation* Dept. Statistics, Stanford Univ.
- [21] Kabashima Y, Wadayama T and Tanaka T 2009 *J. Stat. Mech.* L09003
- [22] Rangan S, Goyal V and Fletcher A K 2009 in *Y. Bengio et al. (eds.), Advances in Neural Information Processing Systems* **22** 1545
- [23] Guo D, Baron D and Shamai (Shitz) S 2009 *Proceedings of the 47th Annual Allerton Conference on Communication, Control, and Computing* 52
- [24] Sakata A and Kabashima Y 2013 *Europhys. Lett.* **103** 28008
- [25] Donoho D L, Maleki A and Montanari A 2009 *Proc. Nat. Acad. Sci. USA* **106** 18914
- [26] Donoho D L, Maleki A and Montanari A 2011 *IEEE Trans. Inf. Theory* **57** 6920
- [27] Rangan S 2011 in *Proceedings of the 2011 IEEE International Symposium on Information Theory Proceedings (ISIT), Austin, Texas* (IEEE, New York), 2168
- [28] Krzakala F, Mézard M, Sausset F, Sun Y F and Zdeborova L 2012 *Phys. Rev. X* **2** 021005
- [29] Hastie T and Tibshirani R 1990 *Generalized Adaptive Models* (Chapman and Hall, London)
- [30] Stein C 1981 *Annal. Stat.* **9** 1135
- [31] Donoho D and Johnstone I 1995 *J. Amer. Statist. Assoc.* **90** 1200
- [32] Mézard M, Parisi G and Virasoro M 1987 *Spin Glass Theory and Beyond* (World Scientific)
- [33] Nishimori H 2001 *Statistical Physics of Spin Glasses and Information Processing: An Introduction* (Oxford: Oxford University Press)
- [34] Nakanishi-Ohno Y, Obuchi T, Okada M and Kabashima Y arXiv:1510.02189.
- [35] de Almeida J R L and Thouless D J 1978 *J. Phys. A: Math. Gen.* **11** 983
- [36] Zou H and Hastie T 2005 *J. R. Statist. Soc. B* **67** 301

- [37] Efron B, Hastie T, Johnstone I and Tibshirani R 2004 *Annal. Stat.* **32** 407
- [38] Mézard M and Montanari A 2009 *Information, Physics, and Computation* (Oxford: Oxford Press)
- [39] Efron B 1986 *J. Amer. Statist. Assoc.* **81** 461